# The Inference Efficiency (Tokens/Sec) Of Domain-Adapted Baichuan-2 Models On The Factcc Benchmark When Scaled To

Assignee Research

May 30, 2026

**Abstract**

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the inference efficiency (tokens/sec) of domain-adapted Baichuan-2 models on the FactCC benchmark when scaled to different batch sizes. Programming robots is complicated due to the lack of 'plug-and-play' modules for skill acquisition. Virtualizing deployment of deep learning models can facilitate large-scale use/re-use of off-the-shelf functional behaviors. 11 claims were extracted from source literature; 11 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: RILaaS: Robot Inference and Learning as a Service. Research question: What is the inference efficiency (tokens/sec) of domain-adapted Baichuan-2 models on the FactCC benchmark when scaled to different batch sizes?.

## 2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.8/10.

## 3 Results

13 papers retrieved. 11 claims extracted; 11 independently verified. Quality review score: 8.8/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

| Claim | Verified | Confidence |
|---|---|---|
| Programming robots is complicated due to the lack of 'plug-and-play' modules for skill acquisition. | ✓ | 0.24 |
| Virtualizing deployment of deep learning models can facilitate large-scale use/re-use of off-the-shelf functional behavi | ✓ | 0.30 |
| Deploying deep learning models on robots entails real-time, accurate and reliable inference service under varying query | ✓ | 0.37 |
| RILaaS is a novel Robot-Inference-and-Learning-as-a-Service platform for low-latency and secure inference serving of dee | ✓ | 0.38 |
| RILaaS includes low-latency and reliable serving with gRPC under dynamic loads by distributing queries over multiple ser | ✓ | 0.31 |
| RILaaS includes SSH based authentication coupled with SSL/TLS based encryption for security and privacy of the data. | ✓ | 0.25 |
| RILaaS includes a front-end REST API for sharing, monitoring and visualizing performance metrics of the available models | ✓ | 0.24 |
| Experiments were conducted to evaluate the RILaaS platform under varying loads of batch size, number of robots, and vari | ✓ | 0.33 |
| Benchmark applications of object recognition and grasp planning were provided as a service. | ✓ | 0.18 |
| The complexity of load balancing was addressed with a reinforcement learning algorithm that optimizes simulated profiles | ✓ | 0.24 |
| The reinforcement learning algorithm outperformed several baselines including round robin, least connections, and least | ✓ | 0.35 |

## References


- https://www.semanticscholar.org/paper/65099bc5fafc9a988a8aaf4fc9cda86cc03ef3e2
- http://arxiv.org/abs/2507.07101v4
- https://arxiv.org/abs/2603.28708